

Based on car collision data in California for years 2019-2020, get an estimate of when the pandemic started

In [1]: `import pandas as pd`

In [2]: `#Reading the data into a dataframe
#Note-data source not on GitHub

car_collisions=pd.read_csv("Downloads/datascience_interview_data/california_car_collisions.csv")`

In [3]: `#viewing the dataframe
car_collisions.head(5)`

Out[3]:

	case_id	jurisdiction	officer_id	reporting_district	chp_shift	population	county_city_location	special_conditions
0	81715	1941.0	11342	212	5	7	1941.0	
1	726202	3600.0	8945	064	5	5	3612.0	
2	8008498	3604.0	975	NaN	5	6	3604.0	
3	8008502	3000.0	5163	951J2	5	5	3040.0	
4	8008506	1942.0	27792	1541	5	7	1942.0	

5 rows × 9 columns

In [4]: `#Dataframe information
car_collisions.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 577859 entries, 0 to 577858
Data columns (total 74 columns):
 #   Column                                Non-Null Count  Dtype
---  -
 0   case_id                              577859 non-null  int64
 1   jurisdiction                          576588 non-null  float64
 2   officer_id                           577257 non-null  object
 3   reporting_district                   227895 non-null  object
 4   chp_shift                            577859 non-null  int64
 5   population                           577859 non-null  int64
 6   county_city_location                 577858 non-null  float64
 7   special_conditions                   577858 non-null  float64
 8   beat_type                            577858 non-null  float64
 9   chp_beat_type                        577858 non-null  object
10   city_division_lapd                   40545 non-null  object
11   chp_beat_class                       577858 non-null  object
12   beat_number                          538939 non-null  object
13   primary_road                         577858 non-null  object
14   secondary_road                       577857 non-null  object
15   distance                             577858 non-null  float64
16   direction                            428093 non-null  object
17   intersection                         573141 non-null  float64
18   weather_1                           575939 non-null  object
19   weather_2                            19419 non-null  object
20   state_highway_indicator               577618 non-null  float64
21   caltrans_county                      16728 non-null  object
22   caltrans_district                    16728 non-null  float64
23   state_route                          16728 non-null  float64
24   route_suffix                         10 non-null     object
25   postmile_prefix                      3294 non-null  object
26   postmile                             16727 non-null  float64
27   location_type                        16728 non-null  object
28   ramp_intersection                    9608 non-null   float64
29   side_of_highway                      16714 non-null  object
30   tow_away                             567167 non-null float64
31   collision_severity                   577848 non-null  object
32   killed_victims                      577858 non-null  float64
33   injured_victims                     577858 non-null  float64
34   party_count                          577858 non-null  float64
35   primary_collision_factor              576277 non-null  object
36   pcf_violation_code                    34 non-null     object
37   pcf_violation_category                574290 non-null  object
38   pcf_violation                        545249 non-null  float64
39   pcf_violation_subsection              210891 non-null  object
40   hit_and_run                          577858 non-null  object
41   type_of_collision                    573395 non-null  object
42   motor_vehicle_involved_with           575037 non-null  object
43   pedestrian_action                     577534 non-null  object
44   road_surface                          574101 non-null  object
45   road_condition_1                     574518 non-null  object
46   road_condition_2                      3841 non-null   object
47   lighting                             575790 non-null  object
48   control_device                       574775 non-null  object
49   chp_road_type                         577858 non-null  float64
50   pedestrian_collision                  19474 non-null  float64
51   bicycle_collision                     14378 non-null  float64
52   motorcycle_collision                  19132 non-null  float64
53   truck_collision                       33881 non-null  float64
54   not_private_property                  577858 non-null  float64
55   alcohol_involved                      59250 non-null  float64
56   statewide_vehicle_type_at_fault       475425 non-null  object
57   chp_vehicle_type_at_fault              468260 non-null  float64
58   severe_injury_count                   577858 non-null  float64
59   other_visible_injury_count            577858 non-null  float64
60   complaint_of_pain_injury_count        577858 non-null  float64
61   pedestrian_killed_count                577858 non-null  float64
62   pedestrian_injured_count              577858 non-null  float64
63   bicyclist_killed_count                 577858 non-null  float64
64   bicyclist_injured_count                577858 non-null  float64
65   motorcyclist_killed_count              577858 non-null  float64
66   motorcyclist_injured_count             577858 non-null  float64
67   primary_ramp                          344 non-null    object
68   secondary_ramp                        4201 non-null   object
69   latitude                              357449 non-null  float64
70   longitude                              357449 non-null  float64
71   collision_date                         577858 non-null  object
72   collision_time                         572491 non-null  object
73   process_date                          577858 non-null  object
dtypes: float64(35), int64(3), object(36)
memory usage: 326.2+ MB
```

In [5]: `#Consider collision_date column-
#Estimation is that the pandemic started approximately around the time there was a sharp decline in the number of collisions

car_collisions['collision_date']`

Out[5]:

```
0      2020-03-14
1      2020-07-26
2      2019-05-14
3      2019-04-30
4      2019-05-29
...
577854    2020-02-23
577855    2020-02-24
577856    2020-02-23
577857    2020-02-27
577858         NaN
Name: collision_date, Length: 577859, dtype: object
```

In [6]: `#Consider rows which don't have null values
new_collision_data=car_collisions[car_collisions['collision_date'].notna()]`

In [7]: `#View dataframe without null values in the collision_date column
new_collision_data`

Out[7]:

	case_id	jurisdiction	officer_id	reporting_district	chp_shift	population	county_city_location	special_conditions
0	81715	1941.0	11342	212	5	7	1941.0	
1	726202	3600.0	8945	064	5	5	3612.0	
2	8008498	3604.0	975	NaN	5	6	3604.0	
3	8008502	3000.0	5163	951J2	5	5	3040.0	
4	8008506	1942.0	27792	1541	5	7	1942.0	
...
577853	91200523	9265.0	022134	NaN	3	9	3900.0	
577854	91200524	9720.0	021806	NaN	3	9	4400.0	
577855	91200525	9855.0	014980	NaN	1	5	3602.0	
577856	91200526	9840.0	021037	NaN	3	7	3313.0	
577857	91200527	9670.0	014558	NaN	1	7	3001.0	

577858 rows × 9 columns

In [8]: `#Confirming that the new dataframe doesn't have null values in the collision_date column
new_collision_data['collision_date'].isnull().sum()`

Out[8]: 0

In [9]: `#visualize the data and estimate the date based on the sharp decline in the number of collisions
new_collision_data['collision_date'].value_counts().sort_index().plot()`

Out[9]: <AxesSubplot:>

